Journal of Network and Systems Management manuscript No.
(will be inserted by the editor)

Joint In-Network Video Rate Adaptation and Measurement-Based Admission Control: Algorithm Design and Evaluation

Steven Latré · Filip De Turck

the date of receipt and acceptance should be inserted later

Abstract The important new revenue opportunities that multimedia services offer to network and service providers come with important management challenges. For providers, it is important to control the video quality that is offered and perceived by the user, typically known as the Quality of Experience (QoE). Both admission control and scalable video coding techniques can control the QoE by blocking connections or adapting the video rate but influence each other's performance. In this article, we propose an in-network video rate adaptation mechanism that enables a provider to define a policy on how the video rate adaptation should be performed to maximize the provider's objective (e.g., a maximization of revenue or QoE). We discuss the need for a close interaction of the video rate adaptation algorithm with a measurement based admission control system, allowing to effectively orchestrate both algorithms and timely switch from video rate adaptation to the blocking of connections. We propose two different rate adaptation decision algorithms that calculate which videos need to be adapted: an optimal one in terms of the provider's policy and a heuristic based on the utility of each connection. Through an extensive performance evaluation, we show the impact of both algorithms on the rate adaptation, network utilisation and the stability of the video rate adaptation. We show that both algorithms outperform other configurations with at least 10%. Moreover, we show that the proposed heuristic is about 500 times faster than the optimal algorithm and experiences only a performance drop of approximately 2%, given the investigated video delivery scenario.

Keywords
adaptive video streaming, multimedia management, IPTV, Pre-Congestion Notification

S. Latré, F. De Turck
Ghent University - Department of Information Technology - iMinds,
Gaston Crommenlaan 8/201, B-9050 Gent, Belgium
Tel.: +32-9-3314988 - Fax: +32-9-3314899 - E-mail: steven.latre@intec.ugent.be

1 Introduction

With the recent evolution towards higher resolution videos (e.g., Full High Definition (FullHD) videos), multimedia services are the biggest services in terms of bandwidth consumption but also the services with one of the highest quality demands. According to [1], video is already the dominant traffic in the Internet and will reach a share of 50% by the end of 2012. On the other hand, video has very high Quality of Experience (QoE) requirements: a lack of resources immediately leads to visual artifacts and a deterioration of the QoE. Protecting existing video services against a loss in available resources is thus an important aspect of optimizing the video's QoE.

The challenge of protecting the resources of existing sessions is not a new one. Several standardisation bodies such as the Intserv framework [2] and TISPAN [3] have proposed admission control mechanisms for managed network environments. Whenever a new video session is requested, the request is sent to a Resource Admission Control (RAC) mechanism, describing the traffic characteristics of the video associated with the session. This RAC mechanism then checks if every router along the path has enough resources to support the new video session, after which the RAC mechanism decides to admit or block the session depending on the state of each router.

While RAC mechanisms have been applied to protect video sessions in the past, traditional admission control mechanisms under perform for two reasons. First, the complexity of the traffic patterns of videos hinders an accurate description of the required resources. Video sessions are known to have a bursty bitrate. Therefore, traditional RAC mechanisms often dimension the required resources on the video's peak rate. Although this successfully avoids any QoE degradation, this is a gross over-dimensioning of the network leading to a loss in network utilization and consequently in a loss of revenue for the operator. More recently, measurement based admission control (MBAC) mechanisms have been proposed that take the admission decision based on local measurements in the network and rely on statistical multiplexing to improve the network utilisation. An example of such an MBAC mechanism is the Pre-Congestion Notification (PCN) mechanism, recently standardized within the IETF [4]. Second, the default decision of a RAC mechanism, admitting or blocking the session, is not always the best option when the requested service is a video. Specifically for video services, other reactions to a scarcity of resources are possible such as offering the video at a reduced video quality. This can be supported by using the Scalable Video Coding (SVC) codec [5], which is a video compressing standard that encodes video into multiple quality layers: a video can be reduced in quality by simply dropping a layer. While reducing the video quality is thus possible, the network provider, managing the network, still needs to determine when to switch to which quality level. To the authors knowledge, this article is the first that combines video rate adaptation with an MBAC system to ensure a smooth video delivery and allows specifying detailed rate adaptation policies.

In this article, we present a joint admission control and video rate adaptation system for SVC-based Video on Demand sessions. The system features a tight interaction with an MBAC system and uses policies to steer the video rate adaptation process, which is responsible for determining which SVC quality layers to drop. Compared to other rate adaptation algorithms such as the suite of HTTP adaptive streaming protocols (e.g., Apple Live Streaming [6], Microsoft Smooth Streaming [7]) the decision on which video quality level to adapt to is not taken by the clients but is performed distributively in the network and controlled by policies, which are defined by the network provider. As such, the video rate adaptation mechanism is particularly useful in a managed IPTV

Fig. 1: Overview of the dynamic video rate adaptation algorithm. Based on the network provider's policy the video rate of existing sessions can be altered as an alternative to blocking new requests. The decision and scaling is executed in the network elements (as detailed in Figure 2).

scenario where a network or service provider wants to control the QoE of the services that are offered to its customers, e.g., to provide QoE guarantees.

Consequently, the contributions of this article are three-fold. First, we integrate an existing standardized MBAC system with a novel video rate adaptation mechanism. We argue that a close interaction between the video rate adaptation mechanism and MBAC mechanism is needed to optimize the QoE. As illustrated in Figure 1, the mechanism not only blocks new requests but also performs a dynamic graceful degradation of existing video sessions through the reduction of the video quality, allowing to make room for new video sessions. Second, we present two different video rate adaptation algorithms that are both able to steer the rate adaptation decision (i.e., which videos are adapted) and focus particularly on the maximization of a network provider's policy under a changing network load. The first one is based on a Linear Programming (LP) model that finds the optimal parameters that maximize the policy, while the second one is a heuristic that calculates, for each connection, the utility of each quality level, which can be seen as the gain that can be obtained by switching to that quality level, and then maximizes the overall utility. Third, we investigate the performance of the approach by evaluating the impact of the algorithm on the obtained QoE, discussing the integration of the MBAC and video rate adaptation mechanism and evaluating the scalability of both decision algorithms.

The remainder of this article is structured as follows. Section 2 provides an overview of traditional admission control mechanisms as well as video rate adaptation mechanisms and their combination. In Section 3, we provide an overview of the video rate adaptation architecture, combining MBAC and video rate adaptation decision algorithm. Section 4 discusses the used MBAC system and the integration with a rate

adaptation decision algorithm (in Section 4.3), which is one of the contributions of this article. In Section 5, two rate adaptation decision algorithms that use the MBAC information are proposed. The complete video rate adaptation system is evaluated in Section 6. Finally, Section 7, summarizes the main findings of this article.

2 Related work

2.1 Admission control

In an effort to protect the QoS in the network, network providers often over-provision the network as a low complexity method to ensure that the available network resources do not exceed the required resources. While over-provisioning might be an interesting solution on a short term, it is not always the most cost-effective. In fact, as the popularity of services increases and technologies and user consumption patterns evolve, it is likely that the required resources will outgrow those available. Furthermore, over-provisioning offers little protection against a sudden change in required resources, e.g., triggered by a flash crowd causing a rapid increase in service popularity. To avoid over-provisioning, additional management solutions are required to prevent over-admission of resources.

Especially in multi-service IP networks, admission control mechanisms have been investigated and proposed in standardized network architectures. The RACS layer in the TISPAN architecture [3] foresees a centralized admission control function that allows policing control and resource reservation in access and aggregation networks. In the TISPAN architecture, the A-RACF functional element, responsible for providing admission control, receives requests for QoS resources and uses the QoS information to decide whether or not to block a session. Another centralized admission control approach is proposed in [8], which introduces the concept of a Bandwidth Broker (BB) in a Diffserv domain. Similar to the A-RACF function in the TISPAN architecture, the BB centralizes information concerning network resources and their usage, the topology and policies. When the set-up of a new flow is requested, the BB is signalled out-of-band for an admission decision. Based on the collected information about the complete management domain, the BB can make an informed decision. The downside of these centralized approaches is the lack of scalability and the difficulty of maintaining the knowledge up to date, especially in large and fast changing management environments. Hierarchical approaches have been suggested to tackle this issue but they have the disadvantage of an eventual cost in coordination among BBs and fragmentation of resources [9]. A complete survey of QoS control mechanisms for Next Generation Networks can be found in [10]

One way to alleviate the scalability issues of centralized approaches is investigated by the Intserv architecture [2] where a distributed admission control mechanism is proposed that assumes admission control functions in each node. The Intserv approach requires the use of a traffic descriptor (called traffic specification or TSPEC) to identify the traffic pattern. Furthermore, the Resource Reservation Protocol (RSVP) [11] is typically used as a convenient explicit resource set-up mechanism. However, for some service types, the patterns are hard to define in a traffic descriptor. Video services are a typical example of such service types. Therefore, the TSPEC often only describes the video's peak rate, again leading to an over-dimensioning and loss in network utilization.

Measurement-based admission control mechanisms (MBAC) do not require such a detailed traffic descriptor. Instead the admittance decision is taken based on either active [12] or passive measurements of the current network resources. In both cases, an MBAC system needs to determine the available bandwidth in order to know whether connections need to be blocked. The available bandwidth can be defined as the remaining amount of traffic that can still be sent along a path in the network without leading to congestion [13]. A wide variety of available bandwidth estimation tools have been proposed in the past (e.g., PatChirp [14], BART [15], Forecaster [16]). Furthermore, several studies have provided a comparison of the different available tools, showing that they differ in accuracy and scalability [17,18]. More recently, Thouin *et al.* [19] proposed a probabilistic available bandwidth estimation tool that links the maximum available bandwidth with the probability that the calculated bandwidth can be achieved. A taxonomy of common available bandwidth estimation tools has been presented by Strauss *et al.* [20]. The calculation of the available bandwidth for admission control purposes has been applied to a wide variety of environments. For example, Ergin *et al.* [21] proposed an estimation method, called DCSPT, which is specifically intended for wireless mesh networks. The DSCPT algorithm allows taking into account interference from carrier sensing neighbours, leading to more accurate results. More directly linked with our work, Davy *et al.* [22] exploited the estimation of the available bandwidth to steer an admission control system in an IPTV environment. This was later extended by Meskill *et al.* [23] to include server selection as well. One of the presented admission control algorithms in [22], links the admittance of connections with their expected revenue and only blocks connections with the lowest revenue. In contrast, our solution uses a policy such as the revenue to steer the rate adaptation process, not the admission control process. However, we use revenue as an example of a policy: other policies are also possible. Additionally, our work also focuses on video rate adaptation. In that sense, both solutions are complementary.

The IETF is currently standardizing an MBAC mechanism to protect the resources of inelastic flows in a Diffserv domain called Pre-Congestion Notification (PCN). In the PCN architecture [4], packets are marked, as a way of in-band signalling, when the network load increases. These marked packets are then interpreted at the edges of the network as a sign of imminent congestion, which allows to timely block connections or even perform flow termination. The PCN Working Group currently standardized PCN's metering and marking behaviour [4] as well as a first encoding scheme for marked packets [24]. Several encoding alternatives have been defined [25,26] as well as different possible behaviours at the edge of the PCN domain [27,28]. For more information about PCN's performance and a survey of PCN's algorithms we refer to [29,30]. A general overview of admission control algorithms is provided in [31]. In our work, we extend the PCN architecture to protect the QoE of videos in a managed network, including the differentiation between different video qualities. Specific admission control solutions for IPTV environments have been studied as well. Often, these solutions are combined with QoS provisioning [32,33].

2.2 Video rate adaptation

Video services are one of the few services that can adapt their rate to still offer their service functionality, but at a reduced QoE. This rate adaptation is achieved by varying the video encoding settings which leads to a varying level of detail in the image.

Triggered by the increasing heterogeneity in terms of end user devices (i.e., ranging from small screen smart phones to high resolution TV sets) there is an increasing focus towards video rate adaptation algorithms that allow degrading the video quality if needed. Traditionally, the video rate was adapted at intermediary nodes through a simulcast technique: several versions of the video are sent by the server and on the adaptation node the adaptation consists simply of selecting the desired version out of the set of available versions [34].

As a simulcast approach introduces considerable overhead more advanced video rate adaptation techniques, called HTTP Adaptive Streaming (HAS), are currently being studied. Recently, several solutions have been proposed that allow changing the rate of HTTP-based video sessions dynamically. Several companies have introduced their own HAS solutions, supported by their own video client software, e.g., Microsoft's Silverlight Smooth Streaming [7], Apple's HTTP Live Streaming [6] and Adobe's HTTP Dynamic Streaming [35]. Furthermore, the Moving Pictures Expert Group (MPEG) is standardizing a similar technique called Dynamic Adaptive Streaming over HTTP (DASH) [36]. All these solutions allow splitting an existing video into smaller segments, each having several video quality levels available. The decision on which quality level is chosen is taken by the video client software and is typically based on QoS metrics such as the average throughput. The downside of this approach is that the service provider has less control over the QoE that is offered to its clients. While a HAS approach might target the QoE maximization of each individual video client, from a provider's perspective, other factors are of importance as well. The approach presented in this article focuses on a global control of the QoE levels offered to the clients, in which the network provider can steer the video rate adaptation decision. Compared to HAS techniques, our solution is more suitable in managed network environment, whereas traditional HAS techniques have their merits in an over the top environment.

Recently, the Joint Video Team of the ITU-T VCEG and MPEG has standardized an extension to the widely used video coding standard H.264/AVC called Scalable Video Coding (SVC) [5]. In SVC, the video is encoded in multiple layers and the video's QoE can be adapted on-the fly by dropping enhancement layers from the stream. The SVC standard only specifies how an SVC video can be encoded and decoded but does not make any recommendations on its integration into a video streaming scenario over a network. The authors of [37] discuss the integration in SVC in a real-time streaming environment and present an overview of use cases for applying SVC on a network environment; one use case is the graceful degradation of videos as the network load increases. Moreover, an overview is given of how SVC can be packetized into RTP streams. The authors also argue the need for Media Aware Network Elements (MANEs) that are capable of adapting the SVC stream based on network providers policies. However, they do not present any algorithmic contribution to implement such a MANE. A similar approach can be found in [38] where the integration of SVC into the MPEG-21 Digital Item Adaptation (DIA) framework is discussed. The MPEG-21 DIA framework provides the tools to enable quality adaptation through, amongst others, XML-driven meta data description and the integration onto typical multimedia network devices such as Set-Top Boxes. Similarly, it describes the tools available for performing the actual quality layer adaptation in SVC, but does not discuss how a network provider can decide to which quality layers it should adapt. The algorithm proposed in this article provides an implementation of such a MANE or adaptation node but also discusses the need for combining it with an admission control system. An initial prototype of such a MANE was proposed in [39]. However, in [39] the authors

assume the adaptation for a single connection and that the adaptation decision depends only on an end-to-end bandwidth estimation, which needs to be signalled. Our approach is distributed amongst the several access nodes and provides more flexibility in defining the adaptation policy. Furthermore, our rate adaptation decision algorithm focuses on the adaptation process of multiple clients. As there are multiple ways to come to the same adaptation configuration with the same QoS as output, the problem we investigate has more degrees of freedom and is thus more complex.

While our approach focuses on application layer measures, SVC has been applied on the MAC-layer as well. There, the use of SVC is optimized to specific network environments such as wireless networks [40, 41]. The goal of their adaptation is to achieve the highest possible quality that still achieves the best possible QoS levels (i.e., no packet loss, limited delay). As such, the metrics that are taken into account are more fine grained such as the Round Trip Time of a connection. The techniques discussed in this article are complementary as they focus more on the network provider's policy: as such, it may be possible that a lower quality is streamed because the network provider favours additional connections instead of a higher video quality.

2.3 Combination of admission control and rate adaptation

Combining admission control with a rate adaptation system that controls the throughput at which a connection is allowed to transmit data has mainly been investigated for wireless networks but not applied to video rate adaptation. For example, in [42], Klein et al., present a combination of call admission control and bandwidth adaptation for heterogenous wireless networks. Similarly to our work, they target the maximization of the network utilisation. However, as they do not focus on video sessions, their main focus is on keeping the blocking and dropping rates at acceptably low levels. Similar combinations have also been applied to other multimedia services besides video. Li et al. [43] discuss the design of a quality aware voice streaming framework for wireless sensor networks. Similar to our work, they argue that an interaction is needed between admission control and voice adaptation. However, as they focus on voice services, their adaptation consists of voice compression and data duplication at the edge of the network over a lossy networkAs such, the approach taken is different as it is targeted for a different network environment and therefore reacts to other stimuli (i.e., packet loss instead of an increased network load). Additionally, in [35], the focus is on optimizing the voice quality of each individual user. While this is supported in our solution, we take a more broader approach: the network provider can define its own policy on how the rate adaptation should occur. The maximization of quality can be such a policy, but others may apply as well. As we have shown in [44], the use of video sessions for an MBAC has important consequences for the configuration and algorithmic design. We derived several guidelines for configuring the PCN MBAC system for protecting video services. In contrast to [44], this paper focuses on the video rate adaptation algorithms and their performance study. The combination of admission control and video rate adaptation is a less studied field. In [45], Fallah et al. combine admission control with a link adaptation scheme for SVC videos in wireless networks. They show that, for wireless networks, a gain can be achieved before dropping SVC quality layers by adjusting the link adaptation mechanism. Although we focus on access networks, when applied to wireless networks, our work is complementary to theirs.

Our work specifically uses policies to control the video rate adaptation process and admission control system. The merits of policies and their architectural integration in the described standardized admission control systems are discussed in [46]. They argue that policies are needed to decouple the configuration of a particular system, tailored to the network provider, from the actual business logic of the system. The policies we use in our work are mainly focused on the video rate adaptation process but can similarly be integrated. In [47], Argririou et al., provide similar policies to control the admission and rate adaptation of connections in a shared bandwidth channel. Similar to our work, they allow defining policies that control the rate adaptation process. However, our work differs from [47] in two ways: first, the policies discussed in [47] focus on QoS optimization, while our policies are more flexible in the sense that other parameters such as revenue and quality parameters can be taken into account as well. Second, their rate adaptation process does not focus on video services but controls the throughput of connections on a shared channel. Therefore the modeled problem is considerably different and the approach in [47] cannot immediately be mapped to the problem of SVC-based video rate adaptation.

This article builds further upon previous work. In [48], we evaluated the performance of different metering algorithms for the PCN admission control system. Additionally, we also presented a static video quality differentiation algorithm, which was able to decide which quality version of a video to admit. Compared to the dynamic rate adaptation algorithm presented in this article, the static video quality differentiation algorithm could only change the quality at time of admittance. In contrast, in this article, the rate adaptation is dynamic and existing videos can be dynamically and gracefully degraded if the network load increases. As such, both algorithms differ significantly as the latter needs to re-evaluate all existing connections as well. Also, the notion of different policies that control the rate adaptation process is novel in contrast to previous work. In [44], several enhancing components were presented for deploying PCN for protecting video services. Although the main focus of these components was on the optimization of network utilization, one component that was briefly discussed was a dynamic video rate adaptation system, which used so-called utility functions to control the rate adaptation. In this article, we present two novel and more powerful video rate adaptation algorithms. In contrast to the utility function based approach presented in [44], where the policies needed to be defined through mathematical functions with many degrees of freedom, the two algorithms in this article allow defining an operator's policy more straightforward through a single maximization function. As such, both the integration of the MBAC system with a video rate adaptation algorithm and the two video rate adaptation decision functions are novel compared to previous work.

In summary, compared to other work, our work is novel for three main reasons. First, we explicitly combine the video rate adaptation system with an MBAC approach: this ensures that both system's decisions are aligned. Second, we focus on assessing the rate for multiple video connections at once. Third, we believe that there is no overall optimal rate adaptation configuration and that the operator must have a way to control the decision. Through our policy-based approach, this is ensured.

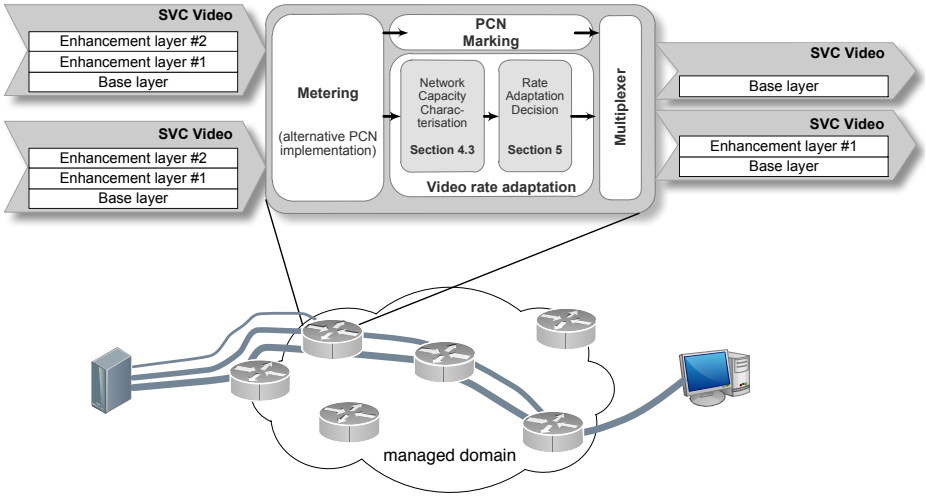


Fig. 2: Overview of the integration of the video rate adaptation algorithm in PCN's architecture.

3 Video Rate Adaptation Architecture

The goal of the video rate adaptation architecture is to dynamically adapt the video quality of existing SVC videos inside distributed MBAC nodes. In our system, we use PCN as MBAC mechanism as it has recently been standardized by the IETF.

Figure 2 illustrates the video rate adaptation architecture and how it is integrated into the original PCN architecture. The video rate adaptation system is deployed on every distributed PCN node. It receives a set of SVC connections as input and dynamically adapts the rate of the existing videos by potentially dropping one or more quality layers (i.e., as part of the video rate adaptation algorithm) and/or marking packets as a sign of a high network load (i.e., as part of the PCN system). The combined video rate adaptation and PCN system works as follows: when a request for a new SVC video arrives, the PCN system is responsible for handling this request. The PCN system can decide to either admit or deny the new SVC connection. When the connection is admitted, this triggers the video rate adaptation decision algorithm. The admittance of a new connection has an impact on the overall network load. The video rate adaptation algorithm can decide to drop one or more quality layers of existing SVC videos or the newly admitted connection. Similarly, when a connection is finished, the video rate adaptation decision algorithm is also triggered. Typically, the video rate adaptation decision algorithm should drop more quality layers as the network load increases. By dropping quality layers, resources become available again and potential new connections can be blocked. An operator typically has many degrees of freedom in tuning the video rate adaptation algorithm including when to perform which quality drop. In our architecture, this is controlled by policies, which is explained in more detail in Section 5.

Similar to the PCN system, the video rate adaptation is distributed amongst the PCN nodes. Every node makes a local assumption of the network status and locally decides whether or not to drop quality layers from an SVC video. As such, it can happen

that a single quality layer is dropped on one node and a second reduction of quality layers occurs further down the path. Each node will drop quality layers to ensure that it can resolve the potential local bottleneck that it experiences. The admission control process is also distributed: each node signals congestion warnings through the marking of packets. However, the admission decision occurs at the edge of the network (i.e., at the ingress node).

As shown in Figure 2, the video rate adaptation algorithm augments PCN's metering and marking functions. In the video rate adaptation process, only when the rate adaptation algorithm decides to stop dropping quality layers, the PCN system should start blocking connections. As such, the rate adaptation algorithm must be aware of the threshold that denotes when the PCN system will start blocking connections. This is calculated in the network capacity characterisation component. We discuss how this threshold can be obtained in Section 4.3, which forms the integration contribution of this article. The calculated threshold is then provided to both the original PCN marking function, responsible for marking packets as an in-band congestion signal, and the actual video rate adaptation algorithm. As the video rate adaptation occurs locally, the rate adaptation algorithm does not require any signalling to other nodes and thus does not require any changes to PCN's marking function. We discuss both components, the integration of the PCN system and the video rate adaptation decision component, in Section 4.3 and Section 5, respectively.

4 Measurement Based Admission Control: The Pre Congestion Notification Mechanism

In this section, we discuss the details of the MBAC mechanism we use in our architecture, being the PCN mechanism, in more detail. Moreover, we detail how the PCN system is integrated into the joint video rate adaptation and PCN system. We discuss only the most important PCN functions, relevant to the video rate adaptation system. For a more extensive discussion on PCN, we refer to [30, 27, 28].

4.1 Original PCN architecture

The goal of the PCN admission control system is to protect the QoS of inelastic flows in a Diffserv domain. Figure 3 provides an overview of the PCN architecture, as standardized in [4]. As illustrated, the PCN architecture defines three node types: a PCN ingress node, a PCN interior node and a PCN egress node. All traffic enters a PCN domain through a PCN ingress node and leaves the domain through PCN egress nodes. Inside a PCN domain (at the PCN interior nodes) and at the PCN ingress nodes, packets are subject to metering and marking. This metering and marking performs a congestion assessment: if the traffic rate is higher than a configured threshold, the incoming packets are marked. When leaving the PCN domain, the PCN egress nodes investigate the amount of marked packets to make an assessment of the congestion. This congestion assessment is then forwarded to the admission control decision point, which may be collocated with the PCN ingress node as illustrated in Figure 3. The decision point calculates a congestion level estimation (CLE) based on the marked traffic rate, reported by the PCN egress nodes. If the CLE is above a configurable threshold,

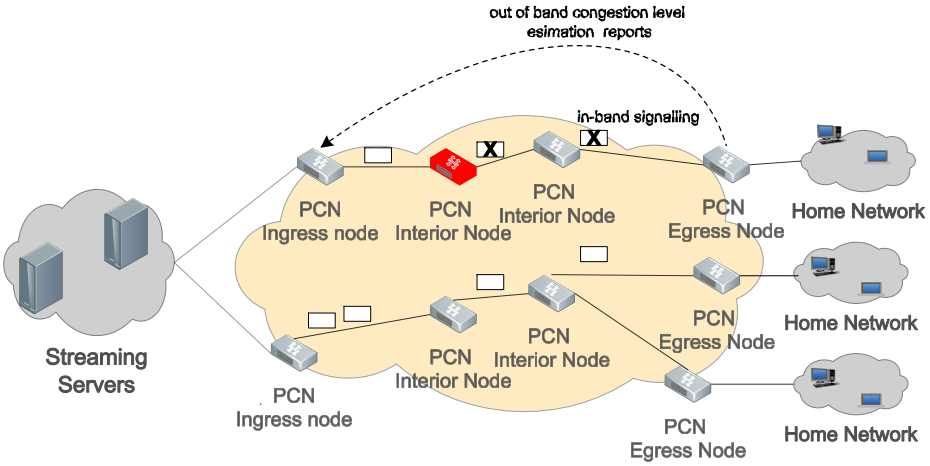


Fig. 3: Original PCN architecture as standardized in [4]. The PCN architecture defines a PCN ingress, interior and egress node

the decision point decides to block all future connections, until new PCN egress reports signal a drop in the CLE value.

The metering and marking function is deployed on the ingress and interior nodes. This function requires the specification of a rate threshold for flow admission and flow termination. For flow admission, an admissible rate AR(l) on each link l of the PCN domain is defined. For flow termination, a sustainable aggregate rate SAR(l) is defined on each link l. By comparing the traffic rate inside the ingress or interior node with these thresholds, the traffic is metered and marked. If the traffic rate exceeds one or both of these thresholds the packets are marked. The marking of packets is done by setting bits in the ECN field of the packet's header. For more information about the PCN encoding options, we refer to [26].

4.2 Modifications to the original PCN metering algorithm: adaptive PCN rate configuration algorithm

The mechanism we use exhibits important modifications to the original PCN metering algorithm, which are necessary to better protect the QoE of video services. In [44], we showed that the bursty nature of video services introduces another important complication with regards to the configuration of PCN's original metering algorithm. The configuration of PCN's configured rate, AR or SAR, does not act as an upper limit on the admitted aggregate bandwidth. Instead, the PCN system will continue to admit connections until all PCN measurements are above this rate threshold. For bursty traffic, PCN's configured rate should be set to a value that allows a certain amount of headroom that is proportional to the variability of the traffic aggregate.

As this traffic aggregate's variability is hard to characterise offline, we use an adaptive rate configuration algorithm, originally presented in previous work in [48], which continuously monitors the variability and sets PCN's configured rate (AR or SAR) accordingly. This adaptive configured rate algorithm has an important impact on the

admittance of connections and video rate adaptation process. The adaptive configured rate corresponds with the threshold at which the PCN system starts blocking connections. These modifications are needed to support the integration with a video rate adaptation algorithm.

4.3 Integration of the PCN system with a video rate adaptation algorithm

In this section, we discuss the novel modifications performed to the PCN system to support the integration with a video rate adaptation algorithm. The proposed rate adaptation algorithm modifies the allowed quality levels as new connections arrive and the load increases. For this, it requires an accurate characterisation of the threshold at which a PCN system starts blocking connections. Therefore, the algorithm uses the output of the PCN metering function (i.e., PCN's configured rate parameter that is continuously adapted). This tight interaction is needed to ensure that both systems are accurately aligned with each other. In other words: that the videos are only being blocked once the video rate adaptation algorithm has lowered the videos to the lowest allowed quality levels. For the algorithm described in Section 4.2, the resulting threshold, denoted by CR, which can correspond with either AR or SAR depending on PCN's configuration, provides a timely but fluctuating assessment of the current threshold limit. In order to ensure a stable output of the video rate adaptation, we first smooth this value by transforming it to a $Limit$ value at time n as illustrated in Equation 1.

$$Limit_n \equiv w \times Limit_{n-1} + (1 - w) \times CR_n \times \theta \tag{1}$$

This smoothing function has two parameters. First, the $Limit$ value is smoothed through an exponentially weighted moving average with weight w to ensure that small oscillations in the configured rate cannot lead to fluctuations in the video rate adaptation decision function. Hence, unlike the CR_n value, the calculated $Limit$ value should be more regarded as an estimation of PCN's threshold on a longer term.

Second, the CR_n value is multiplied by a parameter θ, where $\theta \in [0, 1]$. This θ parameter controls the pro-activeness of the video rate adaptation: if θ is small, $CR_n \times \theta$ will be small as well and the $Limit$ value will result in a more pessimistic assumption of the network's capacity and consequently leading to a quicker adaptation of the video rate. We derive suitable values for both w and θ and show why they are required for the smoothing of the output in Section 6.

The PCN specification [27,28] does not encourage the implementation of other admittance decision algorithms besides either blocking or admitting all connections. However, other MBAC systems may apply more gradual admittance decision algorithms in which only a subset of the future connections is blocked, depending on the network load or because of other parameters such as the expected revenue as proposed by Davy *et al* [22]. In this case, the integration of the video can follow the same principle: based on the integrated MBAC system, the threshold needs to be found that defines when the MBAC system starts blocking the connections (partially). If the MBAC system initially blocks connections partially, the θ factor can be configured higher as both system (i.e., the blocking of connections and the rate adaptation of connections) will coincide with each other.

Table 1: Variables used for the rate adaptation decision function on node n.

Variable	Description
$\mathcal{L}$	The number of outgoing links on node n
l	A specific link on node n
$Limit(l)$	The network's capacity as calculated by Equation 1
$\mathcal{T}$	The number of video types supported by the system.
t	A specific video type
$\mathcal{QL}(t)$	The number of quality levels for type t
q_t	A specific quality level
$\mathcal{C}_{in}(l, q_t)$	The current number of connections of quality level q_t
$B(q_t)$	The expected bitrate of quality level q_t
$Q(q_t)$	The expected QoE score of quality level q_t
$R(q_t)$	The expected revenue of quality level q_t
$\mathcal{C}_{out}(l, q_t)$	The newly calculated number of connections of quality level q_t
$S(l, q_t)$	The maximum allowed share of quality level q_t on link l.

5 Video rate adaptation decision function

5.1 Definition of variables

We first define the problem of the rate adaptation decision on a PCN ingress or interior node n formally. Table 1 summarizes the symbols used for this problem definition. Assume that node n has $\mathcal{L}$ outgoing links, let $l = 1, ..., \mathcal{L}$ denote an arbitrary link on node n. For each link l, there is a calculated limit value $Limit(l)$. Assume there are $\mathcal{T}$ video types present in the network. We define a video type $t = 1, ..., \mathcal{T}$ as a group of videos that can be scaled to the same video quality. Differences in the video types may arise due to differences in encoding settings of the SVC encoder or because the content was delivered by multiple parties. For example, it is possible that there are two video types in the network: one which adapts to two quality levels (e.g., Full HD and SD), and another that allows adapting to three quality levels (e.g., Full HD, HD Ready and SD). Typical VoD providers such as Vudu and Apple often have a handful of video types offered to their customers. Following this definition, each video type t has $\mathcal{QL}(t)$ quality levels. Let $q_t = 1, ..., \mathcal{QL}(t)$ denote an arbitrary quality level of type t. Each quality level will have a dynamic number of active connections ($\mathcal{C}_{in}(l, q_t)$), a static expected mean bitrate ($B(q_t)$), a QoE score ($Q(q_t)$) measured through a visual quality metric and a revenue $R(q_t)$ for offering that particular quality level q_t to the customer. Note that we define the number of active connections of a quality level $\mathcal{C}_{in}(l, q_t)$ as the number of connections that can be served at that quality level, regardless of the previous decision of the video rate adaptation. This means that if the quality level of a particular connection has been changed in the past from q_1 to q_2 by dropping a layer on node n, that connection will still be counted as being part of quality level q_1 because at any given time, the decision function may decide to offer the connection again at quality level q_1 by stopping the dropping of SVC layers of that connection.

The rate adaptation decision function must calculate, for each outgoing link l of every node n and quality level q_t, the number of connections that belong to that particular quality level, denoted by $\mathcal{C}_{out}(l, q_t)$. This is a distributed process and no interaction between the entities is required: a rate adaptation decision function will make the decision independently of other outgoing links or other nodes besides its own. As such, the decision is taken merely based on the information obtained by PCN's local metering function. If multiple bottlenecks occur on the same path, the different

distributed rate adaptation decision functions will each decide to lower the quality of the connections (e.g., potentially deciding to decrease the quality of an already adapted connection). In the remainder of this section, we propose two algorithms for the rate adaptation decision function: both algorithms require a network provider's policy to tune the rate adaptation process. We discuss how this policy can be integrated and modified if desired.

5.2 Linear Programming Formulation

To solve the rate adaptation decision function, a linear programming (LP) model can be defined which finds an optimal solution that maximizes the LP's objective. We abbreviate this algorithm as IVRA$_{LP}$, which stands for In-Network Video Rate Adaptation based on an LP Model.

5.2.1 Decision variables

The model defines $S(l, q_t)$ as the decision variables, which denote the maximum allowed share of quality level q_t on link l. Once the shares $S(l, q_t)$ are calculated the actual video rate adaptation, i.e. the calculation of $\mathcal{C}_{out}(l, q_t)$, is straightforward. Each connection is assigned its highest possible quality level until the share of that quality level is completely saturated. If this is the case, the connection is adapted to the next possible quality level and so on.

Note that not all connections need to be adapted after the calculation of the shares $S(l, q_t)$. Typically, the calculated $S(l, q_t)$ at a given point in time will only slightly differ from the previous calculation. As such, only a subset of the connections need to be adapted in the quality. This can be achieved by intelligently mapping the $S(l, q_t)$ values to the connections in a way that minimizes the number of required adaptations per iteration. For example, suppose we have 9 connections of which 5 of them have been assigned the highest quality and 4 a lower quality. Furthermore, suppose by the next calculation of $S(l, q_t)$ the total number of connections is 10 and the corresponding $S(l, q_t)$ values are 40% and 60% for the higher and lower quality, respectively. Then, it is straightforward to (i) adapt only a single high quality to a lower quality and (ii) assign the new connection the lower quality. As such, only 2 out of 10 connections are adapted.

5.2.2 Objective

The LP's objective corresponds to the network provider's policy and multiple variations are possible based on the details of the policy. An example of a network provider's policy can be to maximize the total revenue of the currently set of offered connections. In this case, the LP's objective is the following:

$$max \sum_{t=1}^{\mathcal{T}} \sum_{q=1}^{\mathcal{QL}(t)} S(l, q) \times R(q) \tag{2}$$

Alternatively, if the network provider's policy is to focus more on the maximization of the QoE, the corresponding LP's objective is the following:

$$max \sum_{t=1}^{\mathcal{T}} \sum_{q=1}^{\mathcal{QL}(t)} S(l,q) \times Q(q) \tag{3}$$

Note that the maximization of the LP's policy only controls the rate adaptation process and not the MBAC system, which is revenue-agnostic. In our approach, the admission control system is only triggered as a last resort mechanism, i.e. when all connections have been adapted to the lowest allowed quality and the only way of protecting the network from congestion is blocking new connections. In this case, all new connections need to be blocked until resources become available again. For other admission control algorithms, which block only a subset of the connections, the admission control algorithm can also be linked with the notion of revenue. This is out of the scope of this paper.

Any policy of which its violation can be quantified as a cost, can be modeled using this approach. As such, more complex policies can be defined as well. For example, a service provider can make a model that represents the quality drop that a client observes as a cost. This quality drop can take into account the subscription level (e.g., gold users require a higher quality than silver users and the cost for a quality drop of a gold user will thus be higher) and device characteristics (e.g., an adaptation to the lowest quality will be less severe for a handheld device compared to a large resolution television screen). Using a weighted combination of these costs, a new policy can be constructed.

Another example of a more complex policy is the differentiation between classes of service. For example, a service provider may choose to map the video streaming of some connections to a best effort service class. A possible policy is then to state that the best effort service class is not allowed to occupy more than X% of the total bandwidth. A violation of this policy (i.e., exceeding the share of bandwidth of the best effort service class by X%) can then be modeled as a cost that increases as the share of bandwidth increases (and is zero if the share is lower than X%). A combination of other policies, focusing on other aspects of the rate adaptation, is possible by making a weighted combination.

5.2.3 Constraints

The constraints of the LP model are the following. First, the total bitrate that is achieved by the video rate adaptation decision must not lead to congestion, hence:

$$\sum_{t=1}^{\mathcal{T}} \sum_{q=1}^{\mathcal{QL}(t)} B(l,q) \times S(l,q) \leq Limit(l) \tag{4}$$

Note that we use the mean bitrate $B(l,q)$ for characterizing the required resources of a quality level and not the peak bitrate. On a long timeframe, the mean bitrate is the best indicator for the required resources. On a shorter timeframe, the required resources may be burstier, but due to the statistical multiplexing of the different connections we can assume that peaks caused by one connection are cancelled out by silent periods of other connections. If, for some pathological cases, an unexpected peak in bitrate

occurs, we can handle this peak for two reasons. First, in Equation 4, we compare with the approximated *Limit* value and not with the admissible rate. This is an important aspect of the approach: by comparing with the smoothed *Limit* threshold we are able to (1) cancel out the required headroom of the variability of the connections and (2) guarantee a level of pro-activeness as defined by the θ parameter to ensure that the video rate adaptation process starts before blocking connections. Second, if the peak in bitrate is so high that the buffer created by the *Limit* parameter is not sufficient, the MBAC system will still avoid congestion by blocking new requests. As the MBAC system measures the traffic aggregate instead of making assumptions on the required resources of every individual connection it is capable of detecting an expected peak in the traffic aggregate.

The second constraint states that all admitted connections need to be taken into account, or in other words, that the total share of connections should be 1:

$$\sum_{t=1}^{\mathcal{T}} \sum_{q=1}^{\mathcal{QL}(t)} S(l,q) = 1 \tag{5}$$

The third set of constraints concerns the differentiation between video types. As it is not possible to adapt the rate between video types, we must ensure that for all video types, the total number of shares that is calculated corresponds with the share of that video type in the total number of connections.

$$\forall t \in \mathcal{T} : \sum_{q=1}^{\mathcal{QL}(t)} S(l,q) = \frac{\sum_{q=1}^{\mathcal{QL}(t)} \mathcal{C}(l,q)}{\sum_{t=1}^{\mathcal{T}} \sum_{q=1}^{\mathcal{QL}(t)} \mathcal{C}(l,q)} \tag{6}$$

Finally, the paradigm of SVC determines that the admitted quality level of a connection can only be the same or lower than the original as the video rate adaptation process works by dropping layers. As such, the last set of constraints states that, for each quality level q_t, the sum of shares of quality level q_t and lower must be lower than the sum of the current share of connections of quality level q_t and lower:

$$\forall t \in \mathcal{T}, \forall q \in \mathcal{QL}(t) : \sum_{i=1}^{q} S(l,i) \leq \frac{\sum_{i=1}^{q} \mathcal{C}(l,i)}{\sum_{t=1}^{\mathcal{T}} \sum_{q=1}^{\mathcal{QL}(t)} \mathcal{C}(l,q)} \tag{7}$$

5.2.4 LP Solution

An optimal solution, which maximizes the objective of the LP, can be computed for the above LP problem using the ILOG CPLEX [49] software package, with the simplex and interior point methods [50]. Note that the above problem is not an Integer Linear Programming (ILP) model as the decision variables $S(l,q)$ are real-valued. LP models can be solved in polynomial time, while ILP models are NP-complete. If the $\mathcal{C}_{out}(l,q_t)$ parameters were used as decision variables, this problem would be transformed to an ILP and thus NP-complete problem.

5.3 IVRA$_{UBH}$: Utility-Based Heuristic

In this section, we present a heuristic that serves as an alternative to the optimal IVRA$_{LP}$. While the IVRA$_{LP}$ approach guarantees optimality in the defined objective,

the model can become large (i.e., more than 1,000 constraints and decision variables) for large scale problems. In such a case, a heuristic is preferred as it can solve the problem more quickly with a limited drop in optimality. Hence, the IVRA$_{LP}$ approach should thus be seen as a benchmark to assess the performance of the heuristic.

To steer the video rate decision process, the heuristic calculates a utility and cost for every connection and every possible video rate adaptation it can take on that connection. We abbreviate this heuristic as IVRA$_{UBH}$, which stands for In-Network Video Rate Adaptation using an Utility-Based Heuristic. IVRA$_{UBH}$ adapts each connections to the quality level that maximises the difference between the calculated utility and cost across the different quality levels. The calculation of utility depends again on the provider's policy. Following the examples of the previous section, if the policy is to maximize the revenue, the corresponding utility of assigning connection c to quality level q can be calculated as the normalisation of possible revenue values:

$$\forall t \in \mathcal{T}, \forall q \in \mathcal{QL}(t) \;\; utility(c,q) = \frac{R(q) - min_{i \in \mathcal{QL}(t)}(R(i))}{max_{i \in \mathcal{QL}(t)}(R(i)) - min_{i \in \mathcal{QL}(t)}(R(i))} \tag{8}$$

Similarly, if the network provider's policy is targeting QoE maximization, the corresponding utility is:

$$\forall t \in \mathcal{T}, \forall q \in \mathcal{QL}(t) \;\; utility(c,q) = \frac{Q(q) - min_{i \in \mathcal{QL}(t)}(Q(i))}{max_{i \in \mathcal{QL}(t)}(Q(i)) - min_{i \in \mathcal{QL}(t)}(Q(i))} \tag{9}$$

The above policies have a similar form as those of $IVRA_{LP}$. Hence, $IVRA_{UBH}$ supports the same complexity of policies as described in Section 5.2.2.

Note that, similar to the $IVRA_{LP}$ algorithm, the above policies relate to the rate adaptation process. We calculate the cost by approximating the average of bitrate cost cb and cost of consecutive switches cs. Preference can be given to the bitrate cost cb or switching cost cs through the weight value w. Hence, the cost $cost(c,q)$ of assigning connection c to quality level q is:

$$cost(c,q) = w \times cb(c,q) + (1-w) \times cs(c,q) \tag{10}$$

The bitrate cost cb is calculated by approximating the negative normalised difference between the connection's bitrate and a value $AvailBW$, which is an estimation of the bitrate available to each connection assuming that all to be adapted connections will receive an equal share of bitrate. This $AvailBW$ will be continuously updated depending on the previous video rate adaptation decisions. If serving the connection at a quality level q where $B(q) \leq AvailBW$ this cost will be zero.

$$cb(c,q) = max(0, \frac{B(q) - availBW)}{availBW}) \tag{11}$$

The switching cost cs is a cost that takes into account previous switching decisions by calculating the number of recent quality switches through an exponentially weighted moving average. If the connection has suffered from various quality switches in the past, this cost will be high. Hence:

$$cs(c,q)_t = w \times cs(c,q)_{t-1} + (1-w) \times S \tag{12}$$

Here, S is 1 if the connection was not adapted to quality level q at time $t-1$. As we want to take into account a limited history window, we set this weight value to 0.9.

Based on these definitions of utility and cost, IVRA$_{UBH}$ is illustrated in Algorithm 1. As shown, IVRA$_{UBH}$ calculates for each connection the utility and cost and selects the quality level that maximises the difference between the calculated utility and cost. Afterwards, the $AvailBW$ parameter is updated to reflect the new situation of bandwidth available for the connections that still need a decision. This update involves subtracting the $AvailBW$ with the bitrate that the new connection will consume. If this is higher than $AvailBW$, the next connections will have less bitrate at their disposal and vice versa.

Algorithm 1 Algorithmic description of IVRA$_{UBH}$

Set $conn$ to total number active connections
$AvailBW \leftarrow \frac{Limit(l)}{conn}$
for all $t \in \mathcal{T}$
  for all $q \in \mathcal{QL}(t)$
    for all $c \in \mathcal{C}_{\setminus\setminus}(l, q_t)$
      $maxUtility \leftarrow 0$
      $levelToScale \leftarrow \phi$
      for all $i \leq q$
        Calculate $utility(c, i)$ and $cost(c, i)$
        if $utility(c, i) - cost(c, i) \geq maxUtility$ **then**
          $maxUtility \leftarrow utility(c, q) - cost(c, q)$
          $levelToScale \leftarrow i$
        end if
      end for
      $C_{out}(l, levelToScale) \leftarrow C_{out}(l, levelToScale) \cup c$
      $conn \leftarrow conn - 1$
      $AvailBW \leftarrow \frac{AvailBW \times (conn+1) - B(levelToScale)}{conn}$
    end for
  end for
end for

6 Performance evaluation results

6.1 Experimental setup

We evaluated the performance of both IVRA$_{LP}$ and IVRA$_{UBH}$ and investigated their interaction with the PCN admission control system. We focused on the maximization of revenues as network provider's policy as described in the previous section, unless stated otherwise.

A VoD scenario was modeled by using an NS-2 based simulator, which is capable of simulating the transmission of real video sequences [51]. As illustrated in Figure 4, a tree-based topology, representing a typical multimedia access network, was used where a video server streams SVC videos to a set of clients. The setup contains one bottleneck where the link capacity decreases from 2 Gbps to 1Gbps. The PCN admission control system was deployed onto this multimedia access network in order to use the network characterization of PCN's metering function. To be more suited for protecting video services the standardized PCN system was adapted with the dynamic rate adaptation algorithm as discussed in Section 4.2. The used PCN implementation uses the single marking mode [28], supporting only flow admission and including the CLE report

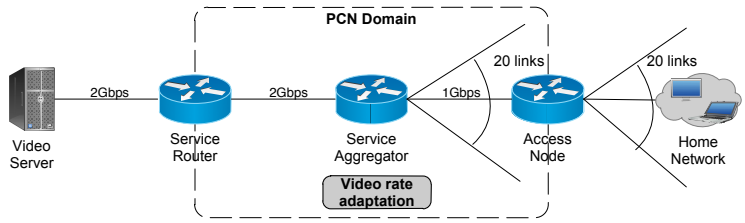


Fig. 4: Used network topology, modeling a typical multimedia access network that offers a VoD service. The video rate adaptation algorithm is deployed on the service aggregator, which forms the bottleneck in the topology.

suppression option with the *T-maxsuppress* timer set to 4.5 seconds and the *CLE-reporting-threshold* set to 0.5. In the IVRA$_{UBH}$ algorithm, no priorities were defined between the switching cost *cs* and bitrate cost *cb*, hence, the weight w was set to 0.5. For the experiments, we focus on the PCN interior node that forms the bottleneck in the network topology. On this PCN interior node, the goal rate was set to 1 Gbps, thus corresponding with the capacity of the outgoing link. Note that the actual configured rate (AR) is varied by the dynamic PCN rate adaptation algorithm. The video rate adaptation algorithms were also deployed on this PCN interior node.

We focused on a scenario with two video types, each with three quality levels: a Full HD video level, an HD ready video level and a Standard Definition video level. Each video item had a length of 90 minutes. Table 2 shows the prices that were used as revenue for each quality level (and both types) together with the mean bit rate and QoE score of each level. A dynamic pricing scheme was used for the evaluated VoD system: users are charged based on the actual quality they receive. Hence, if the quality is adapted to a lower quality, they are charged less and the revenue for the service provider is consequently less was well. Such a dynamic pricing scheme is not yet used in traditional VoD systems. However, there already Content Deliver Network (CDN) provider who charge their customers (i.e., service providers) based on their on-demand consumption (e.g., Amazon's Cloudfront CDN [1] solution). Given the fact that adaptive streaming solutions have only recently been introduced in managed VoD systems, we believe that this is a realistic future pricing scheme.

The experimental setup assumes that it is not possible to switch between qualities of different types: which type was requested was randomly chosen with a uniform distribution. The QoE score denotes the video quality and was characterized using the Structural Similarity Score (SSIM) [52] as video quality metric. The SSIM score is an objective Full-Reference quality metric based upon the assumption that the Human Visual System is more specialized in the extraction of structural information from scenes. The SSIM model takes the original and the distorted signal as input and produces a score between 0 and 1, where 1 stands for perfect quality. The SSIM scores should be interpreted as follows: a video with a SSIM score above 0.9 is indistinguishable from the original, a SSIM score between 0.8 and 0.9 corresponds with a moderate quality while a SSIM score of 0.7 and lower results in a video which is barely watchable. As the SSIM score provides a single value per video frame, we used the mean SSIM score per video as a characterization of the video's QoE. The used prices and bitrates of each

[1] http://aws.amazon.com/cloudfront/pricing/

Table 2: Used quality levels with their corresponding price, average bitrate and QoE score, estimated through the Structural Similarity.

Quality Level	Resolution	Price ($)	Bitrate (Mbps)	SSIM Score
Full HD	1080p	6.00	9.5	0.98
HD Ready	720p	5.00	4.5	0.95
SD	480p	4.00	2.0	0.91

quality level are based on the price models that are currently being used by major VoD providers such as Vudu [53] and Apple [54].

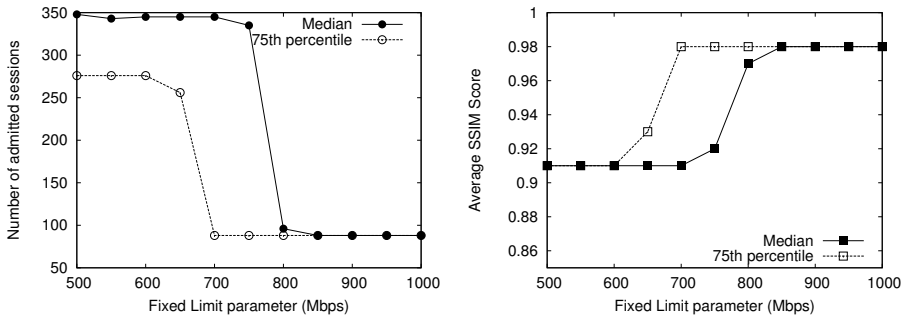
In order to compare IVRA$_{LP}$ and IVRA$_{UBH}$, we performed a one-way ANOVA analysis or t-test on all experiments. An ANOVA analysis is a statistical test that allows determining whether or not the means of two groups of samples is statistically different or if the difference is due to random noise. For our evaluation, ANOVA tests the null hypothesis that the samples obtained through the various metrics extracted from running IVRA$_{LP}$ and IVRA$_{UBH}$ are drawn from the same population. ANOVA provides a decision to reject (i.e., meaning a significant difference between the groups) or accept (i.e., meaning no significant difference) the null hypothesis, given a preconfigured confidence interval.

To model the requests for the SVC videos, we used a production trace of the VoD service of a leading European telecom operator. The simulation time was set to 1 hour and during this timeframe, 1171 videos were requested. The highest request rate observed was 5 requests per second for all clients together. Each experiment was repeated 20 times, the variations between experiments are due to differences in the encoding settings of the videos: various experiments were conducted, each with an alternate encoding of the video content ranging from a set of constant bit rate videos to a set of constant quality videos. We present the average values as well as the calculated confidence intervals, given a confidence level of 99%. In the corresponding figures, the confidence intervals are represented as error bars.

In the remainder of this section, we highlight the need for interacting with a PCN system to measure the network limit and characterize the effect the video rate adaptation functions have on the obtained quality levels. Next, we illustrate the gain of the algorithm by comparing it with a standard PCN system and a video rate adaptation system without integration with PCN. Then, we compare both video rate decision algorithms (IVRA$_{LP}$ and IVRA$_{UBH}$) with each other in terms of the obtained quality level share and the optimality of the algorithms. Furthermore, we investigate the impact of the θ and w parameters of both algorithms. Finally, we investigate the scalability of both algorithms.

6.2 Impact of a fixed configured rate

In this section, we motivate the need for a close interaction between a PCN mechanism and the video rate adaptation algorithm. For this experiment, we fixed the Limit value parameter, which denotes the upper bandwidth limit that can be used for adapting the video rate. Normally, this Limit value is continuously calculated as explained in Section 4.3. As this value is now fixed, the integration between the PCN mechanism and the video rate adaptation is broken in this experiment.



(a) Number of admitted sessions (b) SSIM Score

Fig. 5: Impact of a fixed rate configuration, without any interaction with a PCN system, on the number of admitted sessions and SSIM score.

Figure 5 illustrates the impact of varying the Limit value on the number of admitted sessions (Figure 5a) and average QoE score (Figure 5b) for IVRA$_{LP}$. IVRA$_{UBH}$ provides similar results. As explained above, we investigated different levels of video variability to vary the network experiments; we show the effect for two distinct cases: the median case and the 75th percentile. Figure 5 shows how increasing the Limit value has a decreasing effect on the number of admitted sessions and an increasing effect on the SSIM score. This observation can be explained as follows: as the Limit value is increased, the video rate adaptation algorithm makes a too optimistic estimation of the maximum network capacity. While the theoretical network capacity might be 1 Gbps, in practice the network load measurements will be much lower caused by the burstiness of the videos. Therefore, without any interaction between the PCN system and the rate adaptation decision function, the video rate adaptation algorithm fails to timely respond to a near congestion scenario and hence keeps all admitted videos at the highest quality level. A too high Limit value thus leads to the disabling of the video rate adaptation process.

At the other hand, it is also important to keep the Limit value as high as possible, without disabling the video rate adapation. Although a low Limit value will effectively perform the video rate adaptation, it will make a too pessimistic assumption of the actual network limit and thus drop video layers too soon. This is explained in more detail in Section 6.5.1. There is thus an optimum in selecting the Limit value: the algorithm should use the highest Limit value possible that still performs the actual video rate adaptation and thus has the highest number of admitted sessions. When comparing the median and 75th percentile with each other, we observe that this optimal value changes depending on the variability of the videos: a higher variability (75th percentile) will require a lower Limit value (in this case 650 Mbps) than the median case (where the optimal is 750 Mbps) as the increased burstiness of the video requires a more pessimistic assumption of the actual network limit, and vice versa. These results illustrate an important aspect of the algorithm: without a good estimation of the *Limit* value no optimal rate decision algorithm can be built. Therefore, an integration between the PCN system and video rate adaptation as proposed is thus required.

Table 3: Gain of the algorithm in terms of revenue (instantaneous revenue and aggregated revenue) for various configurations. Our algorithm outperforms all other algorithms and is able to achieve a 10% increase in revenue compared to the second best, corresponding with a not integrated solution.

	Revenue/min after 10 min ($)	Revenue/min after 20 min ($)	Revenue/min after 60 min ($)	Aggregate revenue ($)
No Adapt HD	4.07	5.87	5.87	259.95
No Adapt SD	3.34	5.68	17.73	455.41
Fixed Adapt (LP)	4.07	7.98	14.32	462.36
Fixed Adapt (UBH)	4.07	7.98	14.32	459.62
IVRA$_{LP}$	4.07	8.25	17.73	510.64
IVRA$_{UBH}$	4.07	8.25	17.73	500.75

6.3 Gain of the algorithms

Table 3 illustrates the gain of IVRA$_{LP}$ and IVRA$_{UBH}$ compared to four other configurations: (1) a configuration with only PCN and no video rate adaptation in which all videos are streamed at Full HD quality (which we call 'No Adapt HD'), (2) a similar case but with all videos streamed at SD quality (which we label 'No Adapt SD') and (3) two cases with video rate adaptation enabled (using both the LP model and utility-based heuristic) but without integration between the PCN system and the video rate adaptation mechanism, as investigated in the previous section. As we use the maximization of revenue as a policy to steer the video rate adaptation system, we focus on the obtained revenue as performance metric. We distinguish between two revenue-based metrics: the instantaneous revenue per minute, which indicates the revenue that is generated at that point in time (i.e., by making a weighted combination of the number of admitted connections per quality level) and the aggregated revenue obtained after 1 hour, which can be obtained by summing up the 60 instantaneous revenue values.

We discuss the performance of all six configurations. The 'No Adapt HD' configuration in which all videos are admitted at HD achieves the lowest revenue of only $ 269.95. Indeed, without any video rate adaptation only 88 videos can be admitted and the revenue per video is not high enough to justify the maintaining of every video at the highest possible quality level. However, during the first 10 minutes of the experiment the limited number of active connections allow the 'No Adapt HD' configuration to maximize the revenue. A big aggregated revenue increase can be obtained when all videos are admitted only at SD quality. The aggregated revenue now increases to $ 455.41. Hence, there is a rationale for dropping quality layers to increase the revenue. However, as we can see in the 2nd and 3rd column, without any video rate adaptation we severely lose revenue in the first minutes of simulation as the few connections that are admitted at that time could easily have been admitted at a higher quality. In the first 10 minutes of the experiment we see that the instantaneous revenue of the 'No Adapt SD' configuration ($ 3.34) is considerably lower than that of the other configurations ($ 4.07). Enabling the video rate adaptation (i.e., the last four configurations) allows solving this issue: as the videos are dynamically adapted they can first be admitted at Full HD (thus maximizing the revenue by favoring the highest quality in the first 10 minutes) and later adapted to SD (now maximizing the revenue by favoring a high number of connections).

When comparing the 'Fixed Adapt' configurations and the proposed IVRA$_{LP}$ and IVRA$_{UBH}$ algorithms, we see the need for integrating a PCN and video rate adaptation mechanism in terms of revenue as well. In the 'Fixed Adapt' configurations, the videos were too quickly adapted to lower qualities. This can be seen in the instantaneous revenue values after 20 minutes: an increase in revenue can be obtained compared to the 'No Adapt' configurations but with integration the increase in revenue is considerably higher. Additionally, without integration less connections are admitted at the end of the experiment as not all connections were successfully downgraded to SD. Therefore, the instantaneous revenue values at the end of the experiment (after 60 minutes) are lower as well. By enabling the integration, a considerable increase can be obtained. Our proposed system outperforms all other configurations in all situations and is able to achieve a 10.44% increase in revenue compared to the 'Fixed Adapt' configurations, which does not have the proposed integration.

When comparing both rate adaptation decision algorithms, we see that the optimal IVRA$_{LP}$ algorithm obviously outperforms the IVRA$_{UBH}$ heuristic but that the differences are limited. In the three instantaneous snapshots taken at 10 minutes, 20 minutes and 60 minutes, there is no difference in terms of revenue between IVRA$_{LP}$ and IVRA$_{UBH}$. In terms of total aggregate IVRA$_{LP}$ achieves only a 1.98% higher revenue. Hence, there are situations where both algorithms exhibit different behaviour but this does not occur all the time. This is discussed in more detail in the next section.

6.4 Comparison of IVRA$_{LP}$ and IVRA$_{UBH}$

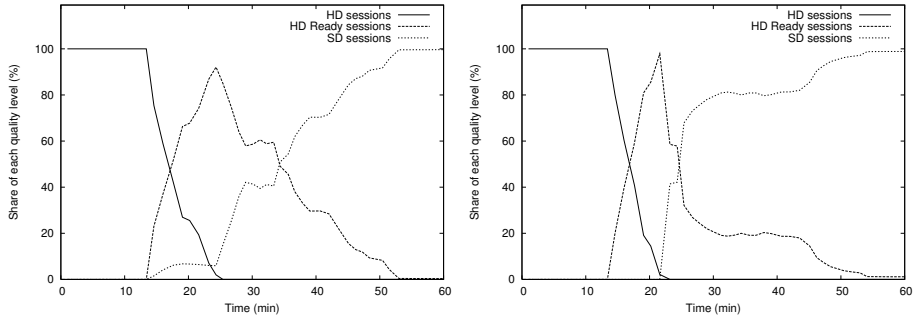
6.4.1 Impact on the quality level share

In order to investigate how IVRA$_{LP}$ and IVRA$_{UBH}$ perform the rate adaptation and to compare their operation with each other, we have characterized the share of each quality level over time for an increasing network load as new requests arrive. Figure 6 illustrates this both for IVRA$_{LP}$ (Figure 6a) and IVRA$_{UBH}$ 6b. Both algorithms are configured with $\theta = 0.7$ and $w = 0.95$. As shown, both algorithms are able to perform a graceful video degradation as the network load increases. Starting out with a non-congested network, all new connections are first admitted at the highest possible quality (i.e., Full HD). As more connections arrive and the network load increases, both existing and new connections are adapted to lower quality levels. Ultimately, the PCN admission control system starts blocking requests for new connections, as all options of video rate adaptation are exhausted.

When comparing both algorithms with each other, we observe that they have similar performance. While there are small differences in when the actual video rate adaptations take place, the share of quality levels they allow at a given time follows a similar behaviour. Especially around the 23 minutes mark and 40 minutes mark we can see some important differences between IVRA$_{LP}$ and IVRA$_{UBH}$: IVRA$_{UBH}$ switches sooner and more drastically to the SD connections than the IVRA$_{LP}$, which features a more smoother transition between HD ready and SD.

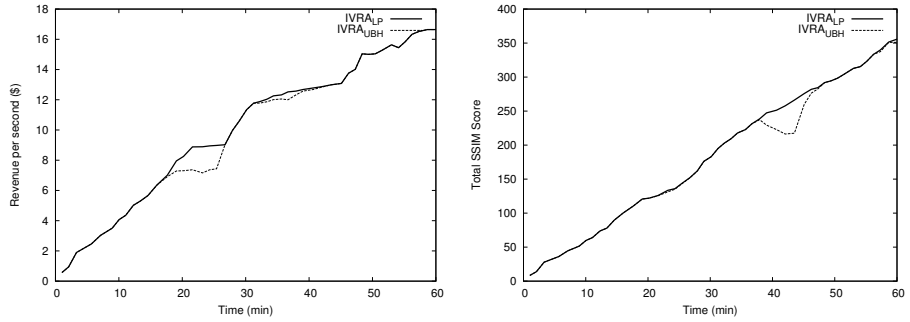
6.4.2 Optimality of IVRA$_{LP}$ and IVRA$_{UBH}$

In this experiment, we compare how optimal both algorithms are in maximizing the specified policy. As IVRA$_{LP}$ is based on an LP model, we know that it will select



(a) IVRA$_{LP}$ (b) IVRA$_{UBH}$

Fig. 6: Impact of both video rate adaptation algorithms (IVRA$_{LP}$ and IVRA$_{UBH}$) over time. As the network load increases and more requests arrive, both algorithms successfully drop to lower quality videos.



(a) Maximization of revenue (b) Maximization of QoE

Fig. 7: Impact of both video rate adaptation algorithms (IVRA$_{LP}$ and IVRA$_{UBH}$) over time on the policy they aim to maximize. Regardless of the policy, IVRA$_{LP}$ outperforms IVRA$_{UBH}$ but the difference is limited.

the video rate configuration that optimizes the policy. IVRA$_{LP}$ is thus by definition optimal in maximizing the configured policy. Therefore, we are particularly interested in the difference in optimality with the IVRA$_{UBH}$ heuristic.

In these experiments, two different optimization policies were configured both for IVRA$_{LP}$ and IVRA$_{UBH}$. The impact of the algorithms on the configured optimization policies was characterized over time. Figure 7 illustrates this impact for a maximization of revenue policy (Figure 7a) and maximization of QoE policy (Figure 7b), respectively. As shown, both policy configurations have similar results. IVRA$_{LP}$ is able to maintain the highest value in revenue or QoE throughout the complete experiment, depending on the configured policy. IVRA$_{UBH}$ often matches the performance of IVRA$_{LP}$ and only experiences a limited performance drop compared to the optimal IVRA$_{LP}$ a few times. Figure 7a shows a performance drop around the 23 minutes mark and 40 minutes mark, which corresponds with the difference in behaviour that was observed

in the quality level share as shown in Figure 6. Similar performance drops can be seen for the maximization of QoE policy as illustrated in Figure 7b. While there are clearly performance drops in the IVRA$_{UBH}$ case, we see that these drops are limited and infrequent. As discussed in Section 6.3, throughout the whole experiment, IVRA$_{LP}$ outperforms IVRA$_{UBH}$ only with 1.98% when the maximization of revenue policy is used. Similarly, the maximization of QoE policy results in a better overall performance of IVRA$_{LP}$ but only with 2.02%.

6.5 Integration of PCN with video rate adaptation

In this section, we investigate the impact of the two parameters, the weight w and θ, that control the integration between the PCN system and the video rate adaptation algorithm.

6.5.1 Impact of the θ parameter

Figure 8 illustrates the impact of an increasing θ parameter on the number of admitted sessions and the underutilisation volume. The underutilisation volume characterizes the average bitrate per second that is not used and is calculated by subtracting the maximum link capacity with the measured throughput as follows:

$$UnderUtilisation = \frac{\sum_{i=1}^{n} \frac{Limit - BW(i)}{mw}}{s}$$

where n is the number of measurements, $BW(i)$ denotes the ith measurement, mw is the time window, $Limit$ is the link capacity and s is the simulation time. Two factors contribute to a non-zero underutilisation volume. First, the burstiness of the aggregate will result in a level of underutilisation: it is therefore not possible to admit connections until the network is completely saturated. Second, specifically for the video rate adaptation algorithm, it is possible that the algorithm decides to lower the video quality too soon. This will result in lower bitrates and thus a higher underutilisation. While the first factor is due to the inherent nature of bursty video, the impact of the second factor can be reduced by tuning the pro-activeness of the algorithm.

As shown in Figure 8, an increased θ value leads to a lower underutilisation volume. This can be explained as follows: the θ parameter controls the pro-activeness of the video rate adaptation. A low θ value will cause the rate adaptation to make a pessimistic estimate of the available network limit and thus results in a lot of connections being lowered in quality too soon, consequently resulting in a higher underutilisation value. Increasing θ thus lowers the underutilisation. There is however a trade-off in increasing θ: as explained before, if θ is too high the bandwidth measurements will never reach the actual calculated limit. This causes the video rate adaptation algorithm to keep most of the videos at the highest quality level, and thus disabling the rate adaptation. Looking at Figure 8, the optimal θ value is the value that still decreases the underutilisation without significantly affecting the number of admitted sessions. In this case, this is 0.7. Across different samples, caused by the streaming of different videos with different encoding settings, the results appear to be stable as well. This is illustrated through the calculated confidence intervals. As shown, the confidence intervals for both the number of admitted sessions and average under utilisation volume are small compared to the observed mean. When comparing IVRA$_{LP}$ and IVRA$_{UBH}$, ANOVA tests showed no significant differences between the two algorithms with a confidence interval of 99.9%.

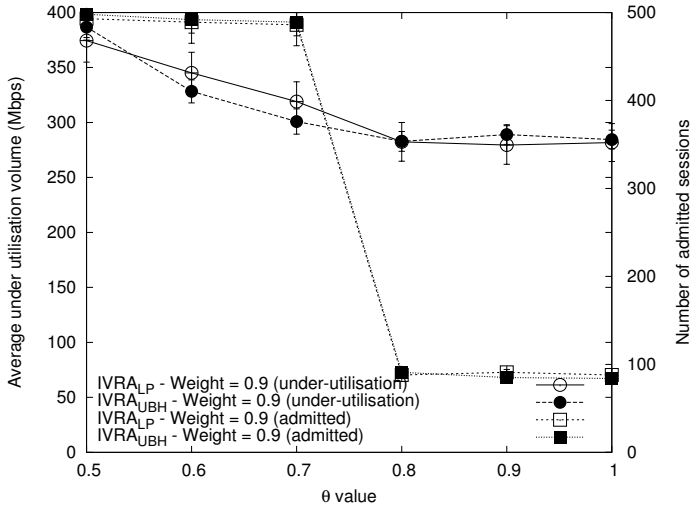


Fig. 8: Impact of an increasing θ on the number of admitted sessions and underutilisation volume.

6.5.2 Impact of the weight value

In the previous sections, we showed how IVRA$_{LP}$ and IVRA$_{UBH}$ are able to perform the desired graceful video quality degradation. Another important requirement for a well performing video rate adaptation algorithm is its stability: a video rate adaptation algorithm should avoid switching back and forth between video quality levels as much as possible. Such oscillations are known to be clearly visible and annoying to the consumer of video services and hence have a destructive effect on the QoE.

Figure 9 illustrates the impact of the weight value on the stability of IVRA$_{LP}$ and IVRA$_{UBH}$. It shows the number of unnecessary quality switches for an increasing weight value. As shown, an increasing weight results in an important reduction of the number of unnecessary quality switches. Figure 9 clearly shows that a weight value of 0.95 or higher is required to avoid instability of the algorithm. This is because the weight value allows smoothing the limit value obtained by the measurement function. When comparing the stability between various values of θ, only a θ configuration of 1.0 does not result in instability. However, in this particular case, the algorithm is stable because there are only a limited amount of video rate adaptations performed as illustrated in Figure 10, which does not reflect the desired behaviour as discussed in the previous section. The confidence intervals show that there is some limited variation across samples, especially when the average number of quality switches is large. However, despite this variation, the confidence intervals are still small enough to obtain stable results, indicating accurate statistical results.

When comparing between IVRA$_{LP}$ and IVRA$_{UBH}$, we see that IVRA$_{UBH}$ has a slightly lower number of unnecessary quality switches than IVRA$_{LP}$. Although IVRA$_{UBH}$ is not optimal in the maximization of the policy, its cost function explicitly takes into account consecutive quality switches. This distinguishing effect is also

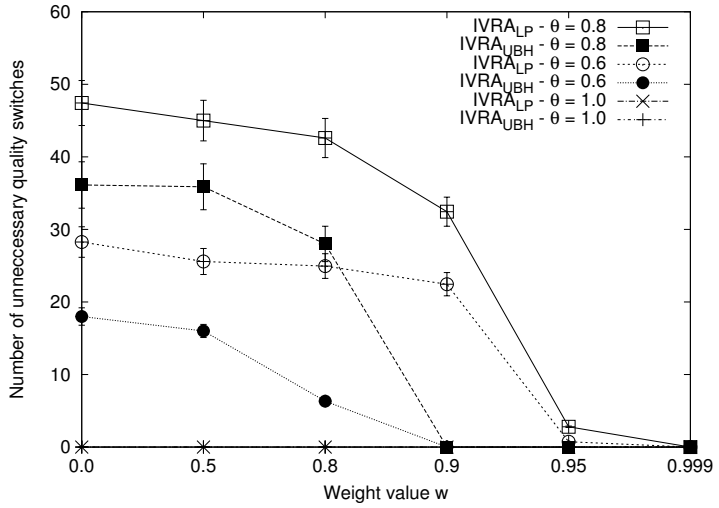


Fig. 9: Impact of an increasing weight value on the number of unnecessary quality switches. The results show that the weight should be at least 0.95 to avoid instability.

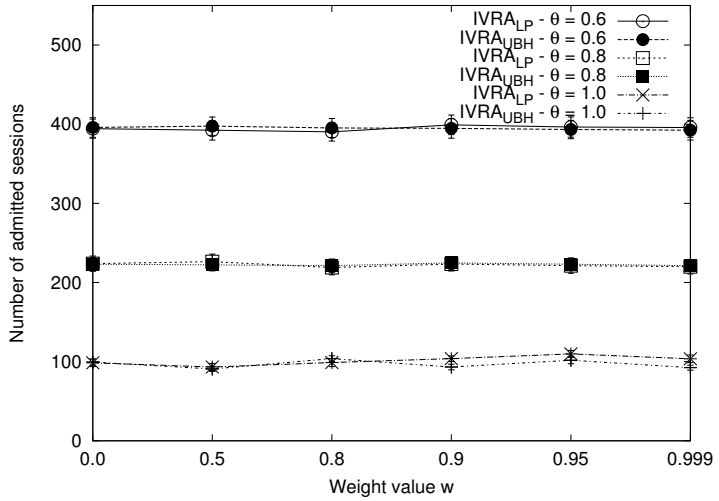


Fig. 10: Impact of an increasing weight value on the number of admitted sessions for three distinct values of θ as defined in equation 1.

justified by an ANOVA analysis: ANOVA tests showed a significant difference between two algorithms for weight values lower than 0.9, with a confidence interval of 99.9%.

While the impact of the weight value on the stability is significant, the weight does not particularly influence the number of admitted sessions. This is shown in Figure 10 which illustrates the impact of an increasing weight value on the number of admitted sessions for three distinct θ configurations. Both algorithms admit approximately the same number of admitted sessions, irrelevant of the weight value. Hence, although the

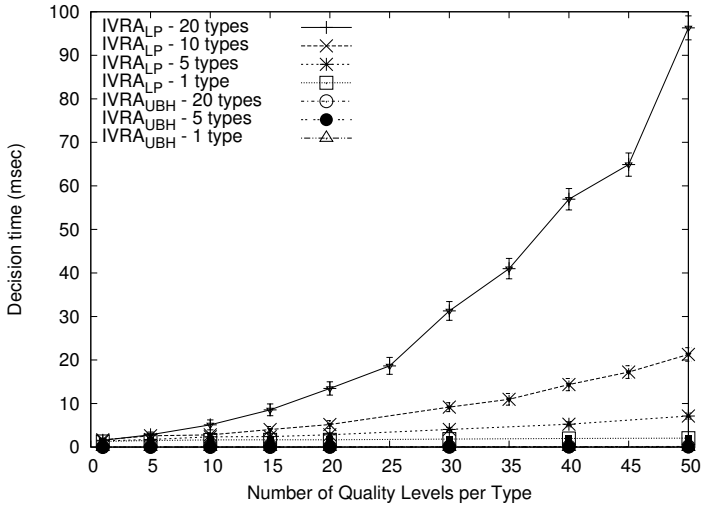


Fig. 11: Time required for calculating the video rate decision function of both IVRA$_{LP}$ and IVRA$_{UBH}$ for an increasing complexity of the problem in terms of number of video types and quality levels.

same number of sessions are admitted, there will be more quality switches to come to this configuration (as illustrated in Figure 9). As shown, the confidence intervals are small and do not change depending on the weight value as well. When comparing both algorithms. an ANOVA analysis showed no significant difference between the two algorithms with a confidence interval of 99.9%.

6.6 Scalability of the algorithms

In this final set of experiments, we investigate the scalability of IVRA$_{LP}$ and IVRA$_{UBH}$ by increasing the problem's complexity. This is done by increasing the number of video types or by increasing the number of quality levels per video type. To evaluate the scalability of both algorithms we characterize the time required to calculate the shares in both algorithms (denoted as the decision time). All experiments were performed on a 1.8GHz Intel Core i7 machine with 4GB of RAM and repeated 1,000 times.

Figure 11 shows the impact of an increasing complexity for both algorithms. As shown, there is a significant difference between both algorithms: IVRA$_{UBH}$ runs much faster than IVRA$_{LP}$. IVRA$_{LP}$ clearly scales polynomially as the number of quality levels per type increases up to decision times which are in the order of tens of milliseconds. This is an issue for high request rates as the rate adaptation function needs to be calculated for each request of a new video and termination of an existing one: if the rate adaptation time is 20 msec or more, the rate adaptation decision alone can only support 50 requests or terminations per second, which might not suffice if a flash crowd occurs.

Figure 12 zooms in on the decision times of IVRA$_{UBH}$. As shown, this algorithm scales linearly with an increasing problem complexity. Moreover, the obtained decision times are in the order of tens of microseconds and thus a 1,000 times smaller than that of

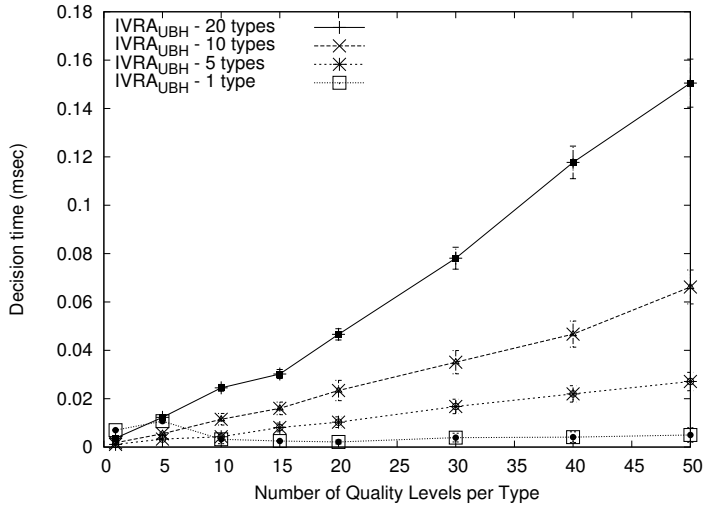


Fig. 12: Time required for calculating the video rate decision function of only IVRA$_{UBH}$ for an increasing complexity of the problem in terms of number of video types and quality levels.

IVRA$_{LP}$. This is an important distinguishing factor between the two algorithms: while the previous results showed no significant difference between IVRA$_{LP}$ and IVRA$_{UBH}$, these results show that IVRA$_{UBH}$ scales considerably better (i.e., 500 times faster), while being still very close in performance in terms of number of admitted sessions, stability and network utilisation to the optimal solution, as calculated by IVRA$_{LP}$. Regarding the statistical accuracy of the results, the obtained confidence intervals, given a confidence level of 99%, are clearly small. This shows that we can have a high confidence in the accuracy of the scalability results. As such, we can conclude there is a clear and significant difference between the scalability of IVRA$_{LP}$ and IVRA$_{UBH}$.

7 Conclusions

In this article, we proposed a joint MBAC and video rate adaptation algorithm for SVC videos in a VoD environment. The recently standardized IETF PCN admission control system was used as MBAC system. The algorithm allows a network provider to specify policies on how existing videos need to be adapted as a function of the network load. These policies therefore allow controlling the overall video rate adaptation process in the network. We discussed two distinct type of policies: the maximization of revenue and the maximization of QoE. However, the algorithm is generic and a network provider can choose to specify its own policies. The system allows a controlled and graceful video degradation before starting to block connections. We argue that the combination and interaction of both mechanisms is required to form an integrated system that allows protecting the QoE of videos in which the two management actions are aligned with each other.Furthermore, we focused on the video rate adaptation decision function that calculates the assignment of connections to quality levels by potentially dropping layers. We presented two algorithms for the rate adaptation decision func-

tion: one LP-based model (IVRA$_{LP}$) that maximizes the provider's policy and one heuristic (IVRA$_{UBH}$) that makes an approximation by estimating the utility of each video adaptation decision. Comparing both approaches with each other, the IVRA$_{UBH}$ heuristic can be seen as a more scalable solution than the optimal IVRA$_{LP}$ algorithm for large scale problems, while IVRA$_{LP}$ serves as a benchmark to characterize how close to optimal the heuristic is able to achieve. Both algorithms allow a provider to change the policy that controls the decision process. Through an extensive simulation-based performance evaluation, we showed that both algorithms are able to accurately control the video rate decision process and are also sufficiently stable in their decision. For example, we showed that the joint PCN and video rate adaptation mechanism is able to outperform a non-integrated combination of PCN and video rate adaptation with 10%, given the investigated network model. Furthermore, in comparing the two algorithms, we showed that the heuristic achieves a good approximation of IVRA$_{LP}$, but has the advantage of an increased scalability. We showed that IVRA$_{UBH}$ scales linearly and consequently requires a factor 500 less computation time for large scale problems compared to IVRA$_{LP}$. In our investigated network model, the increased scalability of the heuristic IVRA$_{UBH}$ comes with a limited cost of 2% compared to the optimal IVRA$_{LP}$ algorithm.

Acknowledgements The research leading to these results has received funding from the European Union's Seventh Framework Programme ([FP7/2007-2013]) under grant agreement n 248775. Steven Latré is funded by grant of the Fund for Scientific Research, Flanders (FWO-V).

References

1. Cisco, "Cisco visual networking index: Forecast and methodology, 2010-2015." White paper.
2. R. Braden, D. Clark, and S. Shenker, "Integrated Services in the Internet Architecture: an Overview." RFC 1633 (Informational), June 1994.
3. ETSI TS 182 019, "Resource and Admission Control Sub-system (RACS); Function Architecture," 2009.
4. P. Eardley, "Pre-Congestion Notification (PCN) Architecture." RFC 5559 (Informational), June 2009.
5. H. Schwarz, D. Marpe, and T. Wieg, "Overview of the scalable video coding extension of the H.264/AVC standard," in *IEEE Transactions on Circuits and Systems for Video Technology In Circuits and Systems for Video Technology*, pp. 1103–1120, 2007.
6. R. Pantos and W. May, "HTTP Live Streaming," Mar. 2012.
7. Microsoft, "Smooth streaming: The official Microsoft IIS site." http://www.iis.net/download/SmoothStreaming - Last accessed on 8 April, 2012.
8. K. Nichols, V. Jacobson, and L. Zhang, "A Two-bit Differentiated Services Architecture for the Internet." RFC 2638 (Informational), July 1999.
9. Z.-L. Zhang, Z. Duan, L. Gao, and Y. T. Hou, "Decoupling QoS control from core routers: a novel bandwidth broker architecture for scalable support of guaranteed services," in *Proceedings of the conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, SIGCOMM '00, (New York, NY, USA), pp. 71–83, ACM, 2000.
10. C. Yun and H. Perros, "QoS control for NGN: A survey of techniques," *Journal Network and Systems Management*, vol. 18, pp. 447–461, December 2010.
11. R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification." RFC 2205 (Proposed Standard), Sept. 1997. Updated by RFCs 2750, 3936, 4495, 5946, 6437.

12. L. Breslau, E. W. Knightly, S. Shenker, I. Stoica, and H. Zhang, "Endpoint admission control: architectural issues and performance," in *Proceedings of the conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, SIGCOMM '00, (New York, NY, USA), pp. 57–69, ACM, 2000.
13. A. Cabellos-Aparicio, F. Garcia, and J. Domingo-Pascual, "A novel available bandwidth estimation and tracking algorithm," in *Network Operations and Management Symposium Workshops, 2008. NOMS Workshops 2008. IEEE*, pp. 87 –94, april 2008.
14. V. J. Ribeiro, R. H. Riedi, R. G. Baraniuk Jiri Navratil, and L. Cottrell, "pathChirp: Efficient available bandwidth estimation for network paths," in *PAM 2003, 4th Passive and Active Measurement Workshop*, (San Diego, CA, USA), NLANR/MNA, UCSD, Apr. 2002.
15. S. Ekelin, M. Nilsson, E. Hartikainen, A. Johnsson, J.-E. Mangs, B. Melander, and M. Bjorkman, "Real-time measurement of end-to-end available bandwidth using kalman filtering," in *Network Operations and Management Symposium, 2006. NOMS 2006. 10th IEEE/IFIP*, pp. 73 –84, april 2006.
16. M. Neginhal, K. Harfoush, and H. Perros, "Measuring bandwidth signatures of network paths," in *Proceedings of the 6th international IFIP-TC6 conference on Ad Hoc and sensor networks, wireless networks, next generation internet*, NETWORKING'07, (Berlin, Heidelberg), pp. 1072–1083, Springer-Verlag, 2007.
17. E. Goldoni and M. Schivi, "End-to-end available bandwidth estimation tools, an experimental comparison," in *Traffic Monitoring and Analysis*, vol. 6003 of *Lecture Notes in Computer Science*, pp. 171–182, 2010.
18. C. D. Guerrero and M. A. Labrador, "On the applicability of available bandwidth estimation techniques and tools," *Comput. Commun.*, vol. 33, pp. 11–22, Jan. 2010.
19. F. Thouin, M. Coates, and M. G. Rabbat, "Large scale probabilistic available bandwidth estimation," *Computer Networks*, vol. 55, pp. 2065–2078, 2011.
20. J. Strauss, D. Katabi, and F. Kaashoek, "A measurement study of available bandwidth estimation tools," in *Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, IMC '03, (New York, NY, USA), pp. 39–44, ACM, 2003.
21. M. A. Ergin, M. Gruteser, L. Luo, D. Raychaudhuri, and H. Liu, "Available bandwidth estimation and admission control for QoS routing in wireless mesh networks," *Comput. Commun.*, vol. 31, pp. 1301–1317, May 2008.
22. A. Davy, D. Botvich, and B. Jennings, "Revenue optimized IPTV admission control using empirical effective bandwidth estimation," *Broadcasting, IEEE Transactions on*, vol. 54, pp. 599 –611, sept. 2008.
23. B. Meskill, A. Davy, and B. Jennings, "Revenue-maximizing server selection and admission control for IPTV content servers using available bandwidth estimates," in *Network Operations and Management Symposium (NOMS), 2012 IEEE*, pp. 319 –326, april 2012.
24. T. Moncaster, B. Briscoe, and M. Menth, "Baseline Encoding and Transport of Pre-Congestion Information." RFC 5696 (Proposed Standard), Nov. 2009. Obsoleted by RFC 6660.
25. G. Karagiannis, K. Chan, T. Moncaster, M. Menth, P. Eardley, and B. Briscoe, "Overview of Pre-Congestion Notification Encoding." RFC 6627 (Informational), July 2012.
26. B. Briscoe, T. Moncaster, and M. Menth, "Encoding Three Pre-Congestion Notification (PCN) States in the IP Header Using a Single Diffserv Codepoint (DSCP)." RFC 6660 (Proposed Standard), July 2012.
27. A. Charny, F. Huang, G. Karagiannis, M. Menth, and T. Taylor, "Pre-Congestion Notification (PCN) Boundary-Node Behavior for the Controlled Load (CL) Mode of Operation." RFC 6661 (Experimental), July 2012.
28. A. Charny, J. Zhang, G. Karagiannis, M. Menth, and T. Taylor, "Pre-Congestion Notification (PCN) Boundary-Node Behavior for the Single Marking (SM) Mode of Operation." RFC 6662 (Experimental), July 2012.
29. M. Menth and F. Lehrieder, "PCN-based measured rate termination," *Computer Networks*, vol. 54, no. 13, pp. 2099 – 2116, 2010.
30. M. Menth, F. Lehrieder, B. Briscoe, P. Eardley, T. Moncaster, J. Babiarz, A. Charny, X. Zhang, T. Taylor, K.-H. Chan, D. Satoh, R. Geib, and G. Karagiannis, "A survey of pcn-based admission control and flow termination," *Communications Surveys Tutorials, IEEE*, vol. 12, pp. 357 –375, quarter 2010.
31. S. Lima, P. Carvalho, and V. Freitas, "Admission control in multiservice IP networks: Architectural issues and trends," *Communications Magazine, IEEE*, vol. 45, pp. 114 – 121, april 2007.

32. L. X. Cai, L. Cai, X. S. Shen, and J. W. Mark, "Resource management and QoS provisioning for IPTV over mmwave-based WPANs with directional antenna," *Mob. Netw. Appl.*, vol. 14, pp. 210–219, Apr. 2009.
33. M. Samie, H. Yeganeh, and M. Shakiba, "A proposed model for QoS provisioning in IMS-based IPTV subsystem," in *Proceedings of the 2009 Fourth International Conference on Systems and Networks Communications*, ICSNC '09, (Washington, DC, USA), pp. 113–118, IEEE Computer Society, 2009.
34. C. Bouras, A. Gkamas, and G. Kioumourtzis, "Performance evaluation of simulcast vs. layered multicasting over best-effort networks," in *Proceedings of the 17th international conference on Software, Telecommunications and Computer Networks*, SoftCOM'09, (Piscataway, NJ, USA), pp. 338–342, IEEE Press, 2009.
35. Adobe, "HTTP dynamic streaming: Flexible delivery of on-demand and live video streaming." http://www.adobe.com/products/httpdynamicstreaming/ - Last accessed on 8 April, 2012.
36. Y. Sánchez de la Fuente, T. Schierl, C. Hellge, T. Wiegand, D. Hong, D. De Vleeschauwer, W. Van Leekwijck, and Y. Le Louédec, "iDASH: improved dynamic adaptive streaming over HTTP using scalable video coding," in *Proceedings of the second annual ACM conference on Multimedia systems*, MMSys '11, (New York, NY, USA), pp. 257–264, ACM, 2011.
37. T. Schierl, C. Hellge, S. Mirta, K. Gruneberg, and T. Wiegand, "Using H.264/AVC-based scalable video coding (SVC) for real time streaming in wireless IP networks," in *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, pp. 3455 –3458, may 2007.
38. M. Wien, R. Cazoulat, A. Graffunder, A. Hutter, and P. Amon, "Real-time system for adaptive video streaming based on svc," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, pp. 1227 –1237, sept. 2007.
39. Y.-M. Hsiao, S.-W. Yeh, J.-S. Chen, and Y.-S. Chu, "A design of bandwidth adaptive multimedia gateway for scalable video coding," in *Circuits and Systems (APCCAS), 2010 IEEE Asia Pacific Conference on*, pp. 160 –163, dec. 2010.
40. C. H. Foh, Y. Zhang, Z. Ni, J. Cai, and K. N. Ngan, "Optimized cross-layer design for scalable video transmission over the IEEE 802.11e networks," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, pp. 1665 –1678, dec. 2007.
41. K. Tappayuthpijarn, T. Stockhammer, and E. Steinbach, "A novel coordinated adaptive video streaming framework for scalable video over mobile networks," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pp. 2893 –2896, sept. 2010.
42. A. Klein, C. Lottermann, C. Mannweiler, J. Schneider, and H. Schotten, "A novel approach for combined joint call admission control and dynamic bandwidth adaptation in heterogeneous wireless networks," in *Next Generation Internet (NGI), 2011 7th EURO-NGI Conference on*, pp. 1 –8, june 2011.
43. L. Li, G. Xin, L. Sun, and Y. Liu, "QVS: Quality-aware voice streaming for wireless sensor networks," in *Distributed Computing Systems, 2009. ICDCS '09. 29th IEEE International Conference on*, pp. 450 –457, june 2009.
44. S. Latré, K. Roobroeck, T. Wauters, and F. De Turck, "Protecting video service quality in multimedia access networks through PCN," *Communications Magazine, IEEE*, vol. 49, pp. 94 –101, december 2011.
45. Y. Fallah, H. Mansour, S. Khan, P. Nasiopoulos, and H. Alnuweiri, "An optimized link adaptation scheme for efficient delivery of scalable H.264 video over IEEE 802.11n," in *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*, pp. 2058 –2061, may 2008.
46. C. Esteve Rothenberg and A. Roos, "A review of policy-based resource and admission control functions in evolving access and next generation networks," *Journal of Networks and Systems Management*, vol. 16, no. 1, pp. 14–45, 2008.
47. N. Argiriou and L. Georgiadis, "A framework for providing user level Quality of Service guarantees in multi-class rate adaptive systems," *Journal of Network and Systems Management*, vol. 16, pp. 375–397, December 2008.
48. S. Latré, B. De Vleeschauwer, W. Van de Meerssche, K. De Schepper, C. Hublet, W. Van Leekwijck, and F. De Turck, "PCN based admission control for autonomic video quality differentiation: Design and evaluation," *Journal of Network and Systems Management*, vol. 19, pp. 32–57, 2011. 10.1007/s10922-010-9183-8.
49. IBM ILOG, "CPLEX 12.0 users manual," in *ILOG Inc., Mountain View, CA*, 2006.

50. G. L. Nemhauser and L. A. Wolsey, *Integer and combinatorial optimization*. New York, NY, USA: Wiley-Interscience, 1988.
51. S. Latré, F. De Turck, B. Dhoedt, and P. Demeester, "Scalable Simulation of QoE Optimization for Multimedia Services over Access Networks," in *The International Conference on Internet Computing (ICOMP)*, 2007.
52. Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, pp. 121–132, February 2004.
53. "VUDU - rent, buy & watch hd movies and tv shows on-demand." http://www.vudu.com.
54. "Apple - itunes - browse the top movie rental downloads." http://www.apple.com/itunes/charts/movie-rentals.

Biographies

Steven Latré obtained a masters degree in computer science in June 2006 and a Ph.D. in Engineering - Computer Science in June 2011, both from Ghent University, Belgium. Since then, he is active as a post-doctoral fellow at the same university. His main research interests are the use of autonomic network management approaches with a special focus on Quality of Experience optimization and federated network virtualization.

Filip De Turck is a full-time professor affiliated with the Department of Information Technology of the Ghent University and iMinds in the area of telecommunication and software engineering. His main research interests include scalable software architectures for telecommunication network and service management, performance evaluation and design of new telecommunication and eHealth services.